UNSUPERVISED TRAINING FOR OVERLAPPING AMBIGUITY RESOLUTION IN WORD SEGMENTATION

ABSTRACT

5   A method for resolving overlapping ambiguity strings in unsegmented languages such as Chinese. The methodology includes segmenting sentences into two possible segmentations and recognizing overlapping ambiguity strings in the

10   sentences. One of the two possible segmentations is selected as a function of probability information. The probability information is derived from unsupervised training data. A method of constructing a knowledge base containing probability information

15   needed to select one of the segmentation is also provided.